Subject: FTB Workflows - OSU
From: Karthik Gopalakrishnan <gopalakk@cse.ohio-state.edu>
Date: Tue, 17 Mar 2009 16:43:35 -0400
To: cifts@googlegroups.com
CC: Dhabaleswar Panda <panda@cse.ohio-state.edu>

Hi All.

Here are the workflows from OSU for FTB-IB.

Summary: FTB-IB is a FTB component that uses the FTB infrastructure to notify other FTB enabled components about failures in the Infiniband Network. Since FTB-IB is a low-level component, it is not really dependant on other components and so does not have to subscribe to events. However, the following two workflow show where FTB-IB could be used.

Workflow 1: Process migration due to network failure

1) An MPI Job is launched on a given set of nodes.

2) Depending on the design of the MPI Library, it could use one or more Ports / HCAs. Assume for the sake of this example that there is only one active port per node.

3) The MPI library subscribes to the FTB_IB_ADAPTER_INFO and FTB_IB_PORT_INFO events. As the name suggests, these FTB Events indicate the status of the InfiniBand Adapters / Ports.

4) Assume that the port that was used by the MPI library goes down. The MPI Library will see a specific failure, maybe ibv_poll_cq() failing with IBV_WC_RETRY_EXC_ERR. However, this information is not sufficient to determine if the failure was due to a port failure on the sender's node or on the receiver's node.

4) The FTB_IB_PORT_INFO event thrown by FTB-IB would indicate to the MPI library (either on the sender or the receiver as the case may be) that a port went down. Armed with this information, the MPI Library can then trigger a process migration from the faulty node to a spare node.

Workflow 2: Port Failover

1) IB Adapters are usually equipped with 2 ports for "High-Availability". A well designed IB network would use both ports of each adapter to ensure that from a given node, all other nodes are reachable through either of the two ports and survive one or more link failures. The same concept could be extended by using multiple IB Adapters per node.

2) In the event of a port failure, the MPI Library can fail-over to

using alternate ports to maintain connectivity.

3) IU Workflow - Interconnect Failure, Section 3.1 talks about this in greater detail.

If other components developers feel they could benefit from other events, apart from those listed in http://nowlab.cse.ohio-state.edu/projects/ftb-ib/software/FTB-IB_Events_1.0.pdf, please let us know. We can add them if it is feasible.

Thanks & Regards,
Karthik